The DAQ System with a RACEway Switch for the PHOBOS Experiment at RHIC

Piotr Kulinich, Pradeep Sarin, and Andrei Sukhanov, *Member, IEEE*

Abstract-- **The PHOBOS data acquisition system based on a RACEway switching network is described. Occupying a single VME crate, the system utilizes 22 PPC750 CPUs working in parallel to compress data from 135,168 silicon pad detectors, and an UltraSPARC VME host for event building and data storage. Lossless Huffman coding is used for compression; this reduces the event size four-fold. The two-host disk array is used to stage data before sending them over Gigabit Ethernet to the RHIC central computing facility. All trigger and control logic is formed using universal programmable logic VME modules, which can be programmed in situ, even when the system is running. The event building and run control software is written using the ROOT framework. The slow control and configuration makes use of an Oracle database to store configuration and monitoring parameters. The system has been taking data from the PHOBOS experiment at RHIC since June 2000. The achieved data-taking rate is 280 events/s or 28 MB/s, with additional disk arrays it can potentially reach 80 MB/s.**

I. Introduction

PHOBOS is one of the four detectors currently installed at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory. Heavy ion collisions at nucleon-nucleon center-of-mass energies of 200 GeV are believed to have produced the highest energy density over an extended volume ever achieved in a laboratory setting. The detector is designed to measure the multiplicity of charged particles produced in these collisions, and to track and identify a small subset of the particles. The setup of the experiment is largely based on silicon pad detectors, which are used in two complementary detector parts. A single layer, large coverage silicon multiplicity detector measures the number of charged particles produced in the collision over a large pseudo-rapidity range. The other detector part comprises two silicon spectrometer arms, which consist of 16 layers of silicon detectors. The spectrometer arms are used to track a fraction (~2%) of the total charged particles emitted in a 2 T magnetic field and to identify these particles by their specific energy loss. Scintillator time-of-flight walls extend the particle identification capabilities to higher momenta in one arm.

A minimum-bias event trigger is provided by a coincidence registered between two sets of 16 scintillator paddle counters, additional trigger conditions are provided by Cherenkov vertex detector and a Zero Degree Calorimeter. Details of the experiment layout can be found in [1].

The data acquisition system (DAQ) at PHOBOS receives data from 135,168 silicon detectors and 1,500 scintillation detectors. It occupies a single VME crate and consists of the following:

- RACEway farm: 22 processing nodes with MPC750, 300 MHz processor, 1 MB of L2 cache SRAM, 32 MB of DRAM; all nodes connected to a RACEway network [2] [3]
- Event builder: Themis USPIIe VME board with UltraSPARC 500 MHz processor, 256 MB DRAM, two 80 MB/s SCSI channels, Gigabit Ethernet NIC, Solaris 8 OS
- Disk array: 600 GB
- Trigger monitor: MVME-2306 board with PPC 604 processor, VxWorks OS
- Data mover: Sun UltraSPARC 3000 server

II. RACEway

The main component of the DAQ is the RACEway network of crossbar switches. It connects terminal elements such as Compute Environment modules (CE), I/O modules and communication bridges. RACEway hardware and software is a product of Mercury Computer Systems, Inc. [4].

The main features of the RACEway and Mercury OS are the following:

- Very high throughput and low latency. Point to point data transfer rate is 160 MB/s. This maximum throughput is easily achievable. When making the connection through the RACEway system each crossbar along the selected path adds only 75 ns to the latency.
- High concurrency and low contention. For example, given four CEs connected to different ports on the same crossbar, CE A can establish a connection with CE B while CE C establishes a connection to CE D with no contention.
- Rapid development. Mercury Systems provides a very rich software environment, specially oriented for real-time computing and parallel processing.

Manuscript received November 5, 2001. This work was supported by the U.S. Department of Energy under Contract DE-FC02-94ER40818 and partially by DE-AC02-98CH10886

P. Kulinich, and P. Sarin, are with the Laboratory for Nuclear Science, MIT, Cambridge, MA 02139-4307 USA

A. Sukhanov is with the Chemistry Department, BNL, Upton, NY 11973-5000 USA (telephone 631-344-8231, e-mail sukhanov@bnl.gov)

- An open and standard interface. RACEway has been approved by ANSI (ANSI/VITA 5.1-1999). To date dozens of companies make RACEway products; there are about 100 products with RACEway on the market at present.

The PHOBOS RACEway hardware consists of VME motherboards, daughtercards and an interlink, they relate to each other as follows:

- Motherboards (6U VME form factor) connect to VME backplanes and have daughtercards mounted on them. They connect the daughtercards to each other and to the RACEway by means of two six-port crossbars (Fig. 1). Each motherboard can carry two daughtercards.
- Interlink-8 (Fig. 2). It plugs into the backplane pins of eight VME P2 connectors. Six six-port crossbar switches connect P2 rows A and C and form the network.
- Daughtercards mount on motherboards. One daughtercard can carry either two MPC750 processing nodes or one I/O device.

III. The PHOBOS DAQ Crate

The RACEway motherboards occupy slots 4 through 11 of the VME crate (Fig. 3). Motherboards in slots 8, 9, 10 and 11 have one processing and one I/O daughtercard (ROUT in slot 8 and 9 and RINT in slot 10 and 11); all other motherboards carry only processing nodes. One of the processing nodes in slot 11 is dedicated as the Master; its main role is to control RACEway I/O devices and to synchronize data transfers. All other boards are Workers.

The signals from the silicon sensors are digitized in Front End Controllers (FECs), which are located in the experimental area [5]. Each sensor is digitized using a 12-bit ADC but samples are stored as 16-bit words. The multiplexed data are transferred to the PHOBOS DAQ crate over two optical fibers using Front Panel Data Port (ANSI approved standard: ANSI/VITA 17-1998) synchronous data flow protocol. Two RACEway input devices (RINT) receive the data under the control of Master CE. It forwards data in a scatter-gather manner to the Workers' local memories. Each Worker watches a buffer in local queue for modification and as soon as it detects arrival of the last word of a dedicated data buffer it starts data processing.

The signals from the scintillating detectors are digitized in a FASTBUS crate; the event builder in slot 3 reads them using a Fast Ethernet link and combines them with the processed silicon data.

The event manager (EMM) in slot 2 is a PCDP VME board (described below); it accepts trigger signals from external trigger modules, generates interrupts for the trigger monitor and distributes a global synchronization marker to all subsystems.

The trigger monitor in slot 1 controls the event manager; it reacts to interrupts from the EMM, provides blocking for the trigger system and monitors the status of the key DAQ components.

IV. Parallel Data Processing

To achieve maximum throughput of the system we use a multicomputer model of process replication with sample set partitioning and circular buffering. The process is simply cloned and each cloned process is given its own processing resources. As a result, this model achieves a linear increase in effective throughput as we add processes. Circular buffering provides additional concurrency by allowing Workers to process data while the Master is writing new data to its local memory.

There are 21 Workers in the system, all running the same code. Each Worker has a local input queue of data frames where it receives data directly from the RINT. The maximal size of a frame is 14 KB; it contains data from two Front End Controllers. The processed events are stored in the output queue located in the Master's local memory. The FEC data block has a global synchronization marker - an 8-bit counter distributed by the event manager at the time when the trigger monitor accepted the event. This marker defines a slot in the output queue where the current data block should be placed; it is also used in the event builder to combine subevents from all subsystems into a single event. The offsets in the slots are controlled by global variables accessible by all Workers and protected by a system spinlock. This scheme provides two levels of parallelism: 1) different Workers can write to the different slots of the output queue, 2) they can write to the same slot but at different offsets (the negotiation for individual offsets is done before the real transfer starts).

The main loop in the Master is as follows:

1. Wait for free space in the input queue.
2. Wait for the end of the conversion in FECs or for the end of the previous transfer, whichever comes first.
3. Start new data transfer to the next buffer in Worker's input queue.
4. Increment input queue write pointer and go to 1.

The main loop in each of the Workers:

1. Wait for a change of the last word of the current local buffer.
2. Check received data for consistency.
3. Compress data.
4. Modify current slot in the output queue and transfer data.
5. If it was last Worker for the current event then increment input queue read pointer.
6. Change to the next local buffer and go to 1

Profiling the working environment shows that the first Worker starts data processing 0.5 ms after the beginning of the first RINT transfer, it takes 1.0 ms to process the frame and the last Worker finishes 1.0 ms after the end of the last RINT transfer. The measured sustained processing performance of the RACEway farm was 400 events/s (40 MB/s); the silicon detectors were not able to provide data at such speed and therefore the system was busy for 50%. Potentially the farm can handle 800 events/s.

V. Huffman Compression

The silicon pad detectors have substantial common mode noise and crosstalk. Signals from empty channels are used in offline analysis for common mode noise correction and for base line shift compensation. Therefore, the data compression is done using the lossless Huffman algorithm [6]. Its main advantage is a high encoding speed and compression ratio close to the Shannon Entropy Limit [7]. Pedestals are subtracted from all channels before compression, resulting in approximately the same mean amplitude over all channels; this makes it possible to use a single coding tree to encode data from all channels. The pedestals and the coding tree are periodically updated and written to the data stream. The typical combined signal distribution is shown on Fig. 4. The entropy (average number of bits per data channel) of the distribution is 4.1. The achieved entropy of the compressed data is 4.5; this corresponds to the compression factor of 3.5.

The RISC architecture of the MPC750 CE is very suitable for Huffman encoding mainly because it has a large set of internal registers (32) and fast L1 cache which is well balanced with a large external L2 cache. The important requirement is that the Huffman table (8 KB in the case of 12-bit data) fits into the 32 KB L1 cache. The measured performance of the compression algorithm is 16 MB/s of input data per Worker and it does not depend on frame size.

VI. Data Moving

The output queue of the Master is accessible from the event builder over the VME backplane. It appeared that the achievable data transfer rate between the RACEway farm and the event builder using VME DMA-D64 block transfer can reach only 37 MB/s. This link was later replaced by a FPDP connection between ROUT in slot 8 and FPDP-PMC interface on the event builder. With this connection the sustained data transfer rate between RACEway and the event builder is 43 MB/s. The main advantage of this connection is that it can be easily duplicated to add more event builders in the system. The event builder runs a multithreaded application, each thread receives data from the connected subsystem and puts the subevent into an output queue according to the synchronization marker; a separate thread writes the complete event to the disk array (RAID level 0). The event builder is able to send data over Gigabit Ethernet directly but to balance the CPU load we stage the data on a disk array. The S100 disk array controller [9] is actually a 2*2 SCSI switch, the first host port is connected to the event builder, and the second is connected to the data mover. Both host ports can run at 80 MB/s. The disk usage is scheduled in such way that while the event builder writes a file to one physical disk the data mover reads another file from another disk and sends data over Gigabit Ethernet to the HPSS storage system at the RHIC central computing facility [10][11]. The disk writing performance is 50 MB/s (CPU is 4% busy) but simultaneous reading from FPDP and writing to the disk saturates at 28 MB/s (CPU is 50% busy) due to a non-optimal driver. The measured sustained throughput from the event builder to HPSS disk cache is 30 MB/s.

VII. Programmable Logic Module

All synchronization and most of the trigger logic on PHOBOS is built using universal programmable logic module – PCDP (Programmable Control and Data Ports) (Fig.5). This is a custom designed A24D16 VME board with front-panel access for 16 differential ECL (dECL) inputs and outputs, and a 17-bit differential TTL bi-directional port (dTTL). It also has a 128 KB FIFO buffer for data transfer and VME interrupt logic. The core of the module is the in-system programmable logic array ispLSI3320 [12], which can be reprogrammed in situ.

The module has 20 ns propagation time from dECL inputs to dECL outputs, and is used to generate first- and second-level triggers, fast clear signals and to scale down input triggers. It is planned to use this module as a fast vertex finder. The dTTL port is used for event synchronization, alarm control and for data transfer. There are 6 PCDP modules in the system. One of them works as the event manager master (EMM), which contains the global event counter. The EMM writes the 8 least significant bits of the counter over dTTL port to dTTL ports of slave PCDP modules in other crates to signal that the new event has been accepted. The slave modules initiate the data taking process in their crates and send a handshaking signal back over a dECL line to EMM when they are ready to accept new events. After receiving the handshake from all slaves, the EMM unblocks the system.

VIII. Data Monitoring

The event builder distributes recent events on demand to the online monitoring computers and to the slow monitoring system. The data mover keeps recent data files available for remote access from event reconstructing computers for prompt full analysis of the selected events. The run control is done using ROOT [8] scripts and a ROOT based GUI. All setup parameters and slow control data are stored in the central Oracle database. The DAQ periodically updates the database with monitoring information, which is stripped from the input data stream.

IX. Summary

A tightly coupled parallel data acquisition system based on the RACEway switching network is described. Each of the 21 computing nodes processes input data using the lossless Huffman compression algorithm with the rate of 14 MB/s producing 4 MB/s of output. Measured sustained performance of the system is 28 MB/s or 280 events/s. With additional disk arrays it can potentially reach 80 MB/s or 800 events/s.

X. Acknowledgments

We wish to thank Sergei Basilev, JINR, Dubna, Russia, who designed all custom boards including EMM and Data Concentrator.

We also would like to thank Burt Holzman, BNL, who developed the online event distributor and managed data sinking to HPSS; Nigel George, BNL, responsible for data validation; Bolek Wyslouch, MIT, for pushing us to integrate the DAQ with the database and maintaining the database integrity and availability at 100% level.

XI. References

[1] B. B. Back, M. D. Baker, D. S. Barton, S. Basilev, R. Baum, R. R. Betts, et al. “The PHOBOS detector at RHIC,” *Nucl. Phys.*, vol. A698, pp 416-419, 2002.

[2] (2001, March). RACE++ Series PowerPC 750 Daughtercards. Mercury Computer Systems, Inc., MA. [Online]. Available: http://www.mc.com/literature/literature_files/ppc750-dc-ds.pdf

[3] B.C. Kuszmaul, (1995, Apr). The RACE Network Architecture. Mercury Computer Systems, Inc., MA. [Online]. Available: http://www.mc.com/literature/literature_files/netwrkarch-wp.pdf

[4] (2001, March). RACEway Interlink Functional Specification. Mercury Computer Systems, Inc., MA. [Online]. Available: http://www.mc.com/literature/literature_files/racewayintrlnk-spec.pdf

[5] P. Kulinich, “Silicon DAQ Based on FPDP and RACEway,” in *Proc of the 6th Workshop on Electronics for LHC Experiments*, 2000, pp 479-482.

[6] D. A. Huffman, “A Method for the Construction of Minimum Redundancy Codes,” *Proc. of the IRE*, Vol. 40, pp. 1098—1101 , 1952.

[7] C. E. Shannon. “A Mathematical Theory of Communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, July, October 1948.

[8] R Brun, F. Rademakers. “ROOT - An Object Oriented Data Analysis Framework” *Nucl. Instr. Meth*., vol. A389, pp 81-86, Sep. 1996. Available: http://root.cern.ch/.

[9] (2000). AVI-S100 Ultra-2 to Ultra-2 SCSI RAID controller. [Online]. Available: http://www.avistor.com/products/controller_products/avi_S100.htm

[10] B. Gibbard, (2000, Feb.). RHIC Computing Facility Processing Systems. Presented at CHEP2000. [Online]. Available: http://chep2000.pd.infn.it/pres/pre_e185.ppt

[11] R. Popescu. (2000, Feb.). PetaByte Storage Facility at RHIC. Presented at CHEP2000. [Online]. Available: http://chep2000.pd.infn.it/pres/pre_c223.ppt

[12] (1999, May). IspLSI 3320 Data Sheet. [Online]. Available: http://www.latticesemi.com/lit/docs/datasheets/cpld/3320.

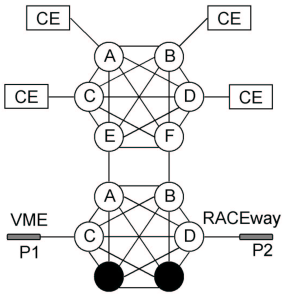

Fig. 1. Block diagram of the RACE VME motherboard showing four Computing Environment elements connected to VME and RACEway.

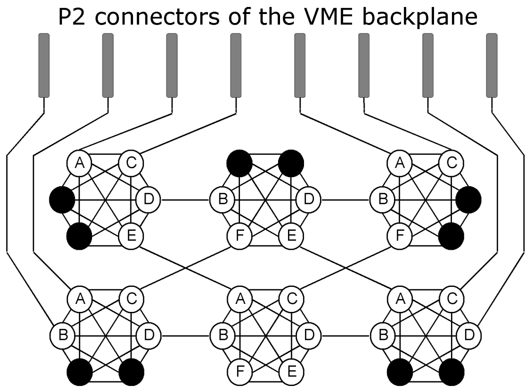

Fig. 2. Eight-slot Interlink.

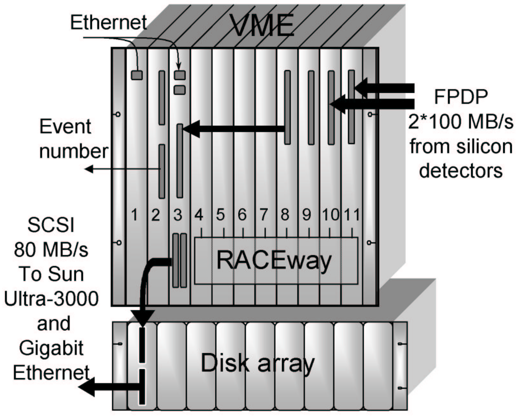

Fig. 3. PHOBOS DAQ crate with disk array. Shown are: RACEway modules in slots 4 through 11, event builder in slot 3, event manager in slot 2 and trigger monitor in slot 1, RINT I/O daughtercards in slots 10 and 11, ROUT I/O daughtercards in slots 8 and 9, FPDP-PMC daughtercard in slot 3.

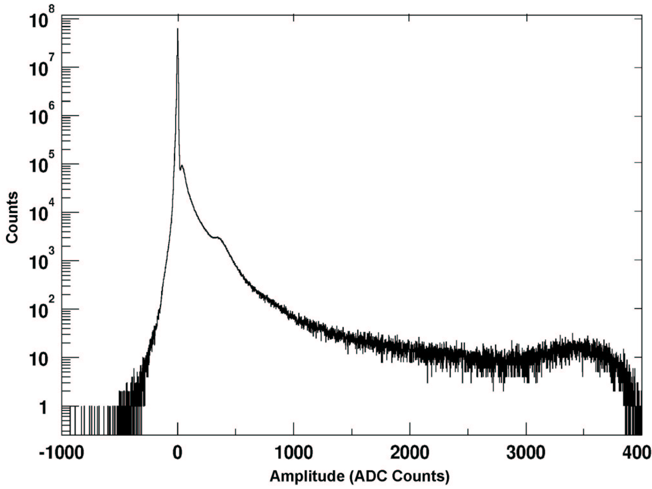

Fig. 4. Histogram of signals from all silicon detectors after pedestal subtraction accumulated for 3000 events. The entropy of this distribution is 4.1.

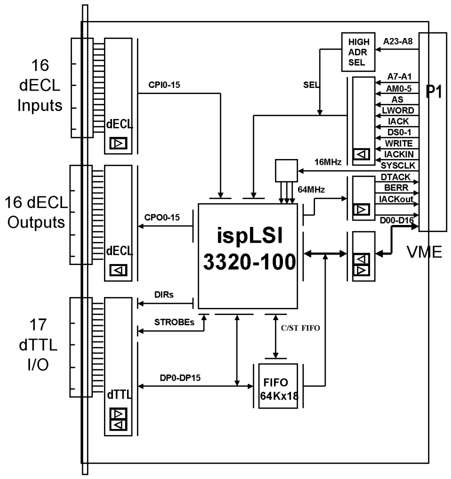

Fig. 5. Programmable Logic Module. The general purpose A24D32 VME board, contains 16 dECL inputs, 16 dECL outputs, 17 bi-directional differential TTL lines, VME interrupts, 128 KB bi-directional FIFO and in-system programmable PLD, connected to all front panel ports. The propagation delay between dECL input and dECL output is 20 ns.